*Scientific Journal of Riga Technical University*
Computer Science. Information Technology and Management Science
*2010*
*Volume44*

# Non-Signature-Based Methods for Anomaly Detection

Pavel Osipov, *Riga Technical University*, Arkady Borisov, *Riga Technical University*

***Abstract* — This paper overviews various approaches to the problem of detecting anomalous behavior within the framework of intrusion detection systems using non-signature-based methods. Each described algorithm has different underlying approach but they all show effective results in the problems of assessing the availability of the wrongfulness of the actions of an authorized user inside an information system.**
**The techniques discussed in the paper use Markov Chains, Hierarchical Hidden Markov Models, algorithms for filtering noise in the signal in the intrusion detection problem, as well as methods based on ontology and agents. Finally, the experimental system developed at Caldas University, Colombia is considered that uses a lot of different approaches aimed to increase anomaly detection efficiency.**

***Keywords* – intrusion detection, statistical model, agents, Markov Models**

## I. Introduction

In today's information society, the task of ensuring information security is one of the most important. There is a whole class of intrusion detection systems (IDS), often using very sophisticated algorithms to detect and prevent intrusion into information systems [1,2, and 3].

In [4] several types of intrusion classification, in particular, by the levels of the analyzed data, are presented. Specifically, four levels are distinguished to analyze the activity for anomaly detection: *network*, *operating system*, *applications and data*. Most of these systems analyze data at all levels of information exchange, from network to application layer. However, the higher the level, the more the models built depend on the specific application. For example, at the network level it is sufficient to analyze the TCP (Transmission Control Protocol) packets, whereas at the application level the context of user interaction with the system has to be taken into account.

In addition to classification by levels of data analysis, IDS can be divided on the basis of analysis of the user activity. There are methods for detection of misuse (Misuse Detection) [5] and abnormal activity (Anomaly Detection). In both cases, illegitimate activities of authorized in target system user are discovered. In the case of misuse detection systems, an approach is used that is based on the patterns of known attacks (Signature-Based systems (SBS)), in the other case –the detection of anomalous activity is based on a statistical model of user behavior (Statistical-Based Intrusion Detection Systems (SBIDS), which is used to analyze user actions.

One of the biggest disadvantages of the SBS approach is the difficulty in detecting new types of attacks, delivering new signatures and updating the model parameters. In turn, the SBIDS approach enables researcher to properly personalize behavior model and analyze not the base of attacks signatures prepared, but the user behavior itself. However, this approach is much more difficult to implement because of the difficulty to formalize it and uncertainty at the stages of adjustment, learning and use.

## II. Problem

This article discusses non-signature-based methods for detection of abnormal activity as more flexible and potentially more effective ones [6]. Most of the methods used in the construction of such systems allow one to evaluate the degree of anomaly of user behavior.

Currently, there are a large variety of methods for constructing the model and evaluating its effectiveness. The main objective of this article is to review the most commonly used approaches.

We consider the following approaches:

- Interpretation of user behavior as a signal, in which case his usual behavior can be filtered out as noise and anomalous considered as signal [9].
- Intrusion detection based on the use of agents [7].
- Ontologies rule in anomaly behavior detection [8].
- Using the Hierarchical Hidden Markov Models (HHMM) [10].
- Construction of user action classifier based on the Markov models [11].
- The prototype system combines several approaches aimed to improve the total efficiency of intrusion detection.

### *A. User behavior as a signal*

Using techniques of time series analysis in the problem of detecting anomalies, normal user behavior is interpreted as a noisebut anomalous as a signal (see Fig. 1). Then, filtering algorithms for signal strength measurements (anomalous behavior) are used. If the signal level is above a certain threshold, then the corresponding action is declared as anomalous.

This approach reduces the number of false positives IDS, as "normal" behavior is filtered out and is not used in the assessment.

The algorithm considered in [9] is a model of user behavior in the form of a Markov Chain (MC). The session of user interaction with the system, consisting of single actions (steps), is not fully analyzed, but only part of it - a sliding window size of $\omega$ symbols (single user transactions in the system). Here, thecomparison with already existing base of

examples of normal behavior is made. For each usertransaction, the minimum Hamming distance for all the steps from the database is calculated.

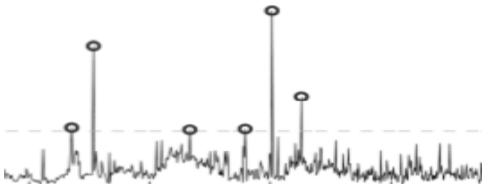

Fig. 1. Anomalous behavior as a signal

The model of user behavior is represented in the form MC $M = (S, p)$, for each step the value of the metric $\alpha \in [0,1]$ is calculated denoting the presence (or absence) of anomalies in the behavior. In this case, the stochastic process model is represented in the form $(H, A)$, where $H$ $\omega$ - vector history of previous steps, $A$ - anomalous values. Then filtering purpose is to remove $A$ from $H$.

At the moment, the possibility of using this approach is being studied for the detection of abnormal activity. The authors plan to conduct the effectiveness and comparative analysis of the approach in the future work.

*B. Hierarchical Hidden Markov Model*

Hidden Markov Model (HMM) [12] is a Markov model, but having a set of unknown parameters. In this case it is required to determine the values of unknown parameters (transition probabilities), based on the known presented variables. Such structures are often used in problems of machine learning, for example, in pattern recognition.

Hierarchical HMM is an extension of the idea of HMM to represent models with a hierarchical structure. HHMM is a structured multi-level stochastic process. HHMM is used for handwriting recognition [13]and visual recognition of action [14].

For a more formal definition of HHMM let us introduce the following terms:

$\sum$ - finite set of states.

$\sum^*$ - all possible combinations $\sum$.

$q_i^d (d \in \{1,..,D\})$ - state with index *i* at the level of *d*.

$|q_i^d|$: - the number of child states, for the root can write $q^d$.

In HHMM, transition between the states at the same level is called horizontal transaction, between different - a vertical transaction.

$A^{q^d} = (a_{ij}^{q^d}) : a_{ij}^{q^d} = P(q_j^{d+1} | q_i^{d+1})$ - probability of horizontal transactions from state *i* to *j* for the subset of nodes $q^d$.

$\prod^{q^d} = \{\pi^{q^d}(q_i^{d+1})\} = \{P(q_i^{d+1} | q^d)\}$ - vector of initial probabilities for a subset $q^d$.

$B^{q^D} = \{b^{q^D}(k)\} : b^{q^D}(k) = P(\sigma_k | q^D)$ - probability that the state $q^D$ gives symbol $\sigma_k \in \sum$.

Based on this, HHMM can be described in the following form:

$$\lambda = \{\lambda^{q^d}\}_{d \in \{1,..,D\}} = \{\{A^{q^d}\}_{d \in \{1,..,D-1\}}, \{\prod^{q^d}\}_{d \in \{1,..,D\}}, \{B^{q^D}\}\}$$

At each level (except the root) there is a final state, after which the process moves to the parent for this subset of the state. This condition allows the use of recursive algorithms on HHMM.

Example topology HHMM is shown in Figure 2.

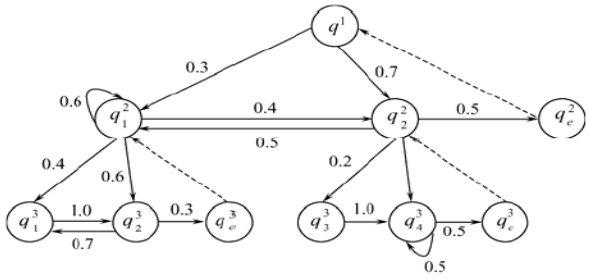


Fig. 2. Topology of a simple HHMM

Detecting anomalies can be interpreted as a problem of the hierarchical nature. Used by user software is distributed by type of functionality, the software themselves are using system calls at different levels.

The most important component of the algorithms using the HMM, is the calculation of the unknown characteristics of the model; for that task the Baum-Welch [13] algorithm is used.

In [10] it is emphasized that as a result of restrictions imposed on HHMM time complexity for calculating the anomalies in the network equals $O(NT^3)$, where *N* – number of states, *T* – number of considered transactions at each step. At the same time, similar, but based on HMM algorithm has the complexity $O(N^2T)$.

As an example of this approach to the problem of intrusion detection [12], the use of three-level HHMM on statistics of real data - system calls of UNIX - server of the University of New Mexico (UNM)is examined.

At the stage of training, the algorithm of Baum-Welch is used for calculating the values of hidden parameters of the model. Then, a base of system calls for analysis is constructed. The third stage is testing when the sliding window test is used on the test data - the analyzer, which returns a sequence of data and calculates for each sequence similarity with the test sequence*O*. The probability of anomaly $P(O | \lambda)$ is calculated depending on the desired threshold.

For the same test data a classifier is created, which uses a simple HMM, and the accuracy of the analysis is compared with the results of HHMM. In order to be able to compare the numerical data, overall response receiver operating characteristic (ROC)was calculated.

Fig. 3 illustrates the accuracy of abnormal activity detection for both methods.

The closer the result to the upper left corner of the chart, the more accurate it is. It is evident that the considered test data method using HHMM showed slightly better results.

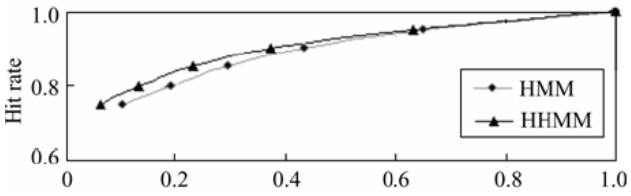


Fig. 3. Comparison of the accuracy of HMM and HHMM

### *C. Ontologies in the problems of anomaly detection*

Most attempts at classifying types of intrusions into information systems are created as a result of the taxonomy of attacks distributed on several grounds. The resulting taxonomy is difficult to use in other systems, other than those for which it was designed. This limitation cannot simply be circumvented by using the taxonomy to store information about the interaction of elements.

To prevent similar developments on the classification of intrusion methods and attributes, more flexible tools, the *ontology* [15] should be used.

Creating the ontology of signs of the intrusion will allow using it in different programs (the separation of logic intrusion detection systems (IDS) and the data model) of machine interaction in automatic mode, that is, the program will be able to use the terms of the current problem domain without a specific setting and attracting experts. It also allows you to create distributed IDS, when a central ontology and queries to itare used.

In [4] an attempt to create such ontology and testing its effectiveness is described.

Ontology is based on the previous studies, whose aim is to classify signs invasions. Also the language for describing ontology, DARPA [16] and tools for working with the constructed model - DAML-JessKB [17] areused.

When creating the considered ontology about 4000 different types of attacks to information systems were analyzed. Also the existing researches were reviewed, and some of them were included as integral parts of the final ontology.

Key attributes of the created ontology are shown in Fig. 4. The main categories of the model are:

- System components (most often attacked). Consists of a stack of network protocols, operating system and applications.
- The essence of the attacks. It consists of the validation errors of entered information, a buffer overflow, error of handling boundary values of such input data and unexpected information.
- Consequences of the attack. As a result, a denial of service attacks, unauthorized access and loss of privacy may occur.
- The position of attack. Separation of the *external*, *local* and *foreign / local.*

Lower levels of abstraction provide a more detailed description of its components. For example, the class Deny of Service has these subsites:Syn Floods, Mailstorms, Pings of Death, and all the other main types of Denial of Service attacks.

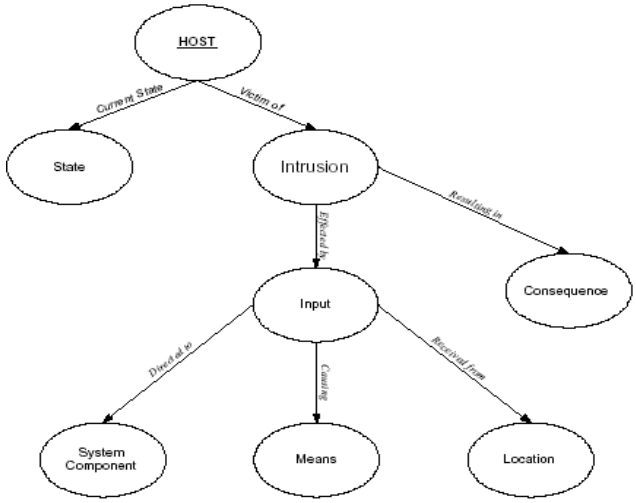


Fig. 4. The upper level of abstraction of ontology

As an example, detection of the famous "Mitnick Attack" [18] is used. It consists of a series of attacks at different levels and cannot be comprehensively interpreted as typical IDS, only components. However, using the central ontology used IDS, producing different levels of monitoring, a sequence of alarm messages can be combined into one specialized rule for exactly this type of attack.

### *D. Agent-based approach to intrusion detection*

In [7], a method for detecting intrusion to an information system based on the use of agents is described. The overall structure of the approach is shown in Fig. 5.

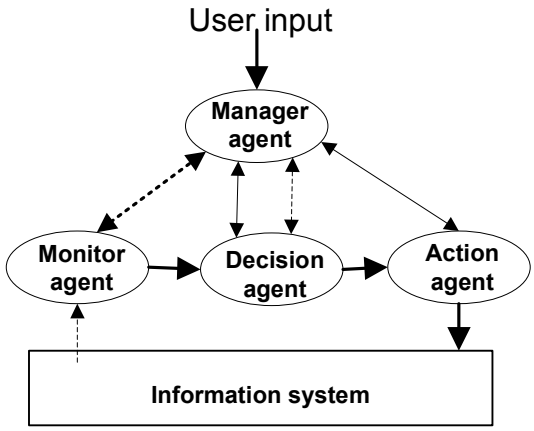


Fig. 5. The overall structure of the interaction of agents

This system is based on open tools for creating multi Wednesday COUGAAR [19]. Using COUGAAR allowed the authors to focus on the core logic of the system, making the technical implementation of agents and protocols for their interactioneasier.

The system is based on the interaction of four agents located at various levels of data exchange and together carrying out the monitoring of current activities.

- Manager agent - is the focal point of action of other agents. Its main task is to manage the tasks and data flows between the other agents. Also, in the case of a distributed

environment, the agent manager communicates with the agent managers in other nodes.

- Decision agent - is responsible for making decisions about the level of the current anomalous activities being analyzed. It contains a variety of analysis modules, such as a module of fuzzy logic, classifiers and a database. The module *fuzzy logic* is used because often the difference between "normal" and "suspicious" behavior has no clear boundaries and methods of fuzzy logic can significantly reduce the number of false positives.
- Action agent - reports on the status of a target system using a specialized language messaging IDS IDMEF (Intrusion Detection Message Exchange Format). In addition, the agent of action gives its recommendations on possible further action (for example, to complete the process, prevent user access to the system, to inform the administration).
- Monitoring agent - collects all information necessary to analyze by decisions agent. Operates at all levels of the analyzed system. To increase the effectiveness of this agent, a specialized module describing the current subject area might be used.

*Example of system usage*

1. The user requests some information, and the manager agent sends it to the monitoring agent for analysis.
2. Analysis agent begins to collect current information from all available levels and analyze it to determine the presence of its deviation from the norm.
3. If any deviation from the norm is detected, the whole context is sent to the decision agent.
4. Decision agent involves its analysis modules, such as fuzzy logic module or other classifiers for identifying the level of the anomalous of the current context.
5. The result of analysis is sent to the action agent, which sends its opinion to the manager agent in the format of the object IDMEF.

The system was successfully tested on several types of test attacks, and showed good results - in some test cases, 100% invasions were found.

### *E. Combining the above methods*

In addition to the methods discussed above there is an interesting example of joint use of various technologies of intellectual data [20] for intrusion detection.

As part of creating experimental IDS OntoIDPSMA (Ontological Intrusion Detection System and Prevention Multi-agent system) developers from Colombia University of Caldas decided to use different technologies at different stages of the analysis of current situation.

The overall structure of OntoIDPSMA is depicted in Fig. 6. All the incoming TCP packets pass through several stages of analysis using different technologies, and as a result, IDS issues an opinion on the possibility of its admission into target information system.

Most analysis is done using the agent-based approach similarto that described above. Agents exchange information in theIDMEF format. To present knowledge about the symptoms of attacks and recommended appropriate responses, ontologyis used. Communicating agents with ontologies were implemented using OWL (Web Ontology Language [21language]).

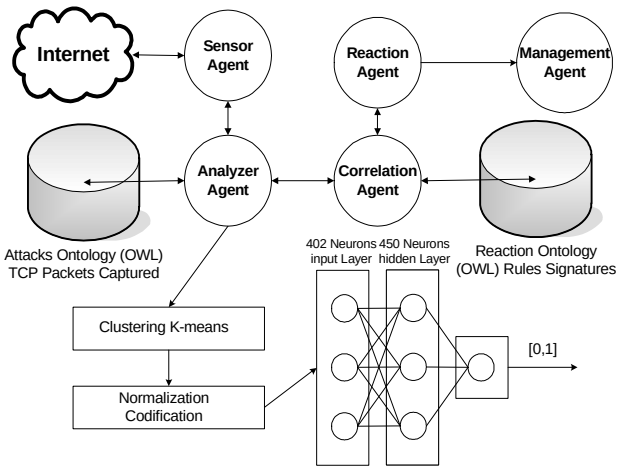


Fig. 6. The overall structureofmodule interaction in OntoIDPSMA

Each type of input data (the headers of packets IP (Internet Protocol), the requested ports, data types and protocols, etc.) were further classified using the method of K-means, the result of classification for all parameters normalized and fed to the input of neural network, which finally gives the result - the presence of anomalies in the analyzed query.

## III. CONCLUSIONS

Nowadays, the increasing rates of growth of information systems make the use of the descriptive approach to intrusion detection even more difficult because up-to-date complex programs have diverse possibilities of impact on the system.

As a result, non-signature-based approach to intrusion detection, which can detect new types of attacks, is becoming even more topical.

Various methods for constructing statistical models to analyze the state of the system and user behavior exist. However, since it is difficult to formalize the problem, an approach that would have a distinct advantage over others and could be used in systems-level enterprise is not set up yet.

In this paper some of these methods are described. Each is significantly different from the others and has both strengths and weaknesses.

## REFERENCES

[1]. **T. F. Lunt**, "Detecting Intruders in Computer Systems," 1993 Conference on Auditing and Computer Technology, SRI International.

[2]. **M. M. Sebring**, **R. A. Whitehurst**, "Expert Systems in Intrusion Detection: A Case Study," The 11th National Computer Security Conference, October, 1988.

[3]. **T. Kohlenberg**, (Ed.), Alder, Raven, Carter, Dr. Everett F. (Skip), Jr., Esler, Joel., Foster, James C., Jonkman Marty, Raffael, and Poor, Mike, "Snort IDS and IPS Toolkit," Syngress, 2007, ISBN 978-1-59749-099-3.

[4]. **J.Undercoffer**, **J.Pinkston**, **A.Joshi**, **T. Finin**. A Target-Centric Ontology for Intrusion Detection // IJCAI-03 Workshop on Ontologies and Distributed Systems, Acapulco, August 2003.
[5]. **C. Warrender**, **S. Forrest**, **B. Pearlmutter**: Detecting Intrusion Using System Calls: Alternative Data Models. IEEE Symposium on Security and Privacy, May (1999)
[6]. **G.J.Mun**, **Y.M.Kim**, **D.K.Kim**, **B.N.Noh**, Network Intrusion Detection Using Statistical Probability Distribution. M. Gavrilova et al. (Eds.): ICCSA 2006, LNCS 3981, pp. 340 – 348, 2006.© Springer-Verlag Berlin Heidelberg 2006
[7]. **D.Dasgupta**, **F.Gonsalez**, **K.Yallapu**, **J.Gomez**, **R.Yarramstetti**. CIDS: An agent-based intrusion detection system. Computers & Security, 2005, 24:387- 398.
[8]. **G.Isaza**, **A.Castillo**, **M.López**, **L.Castillo**. Towards Ontology-Based Intelligent Model for Intrusion Detection and Prevention. "Computational Intelligence in Security for Information Systems", Springer Berlin Heidelberg - Berlin, Heidelberg, 2009.
[9]. **S.Jha**, **L.Kruger**, **T.Kurtz**, **Y.Lee**, **A.Smith**. A Filtering Approach To Anomaly Detection and Masquerade Detection. Technical report, University of Wisconsin, Madison.
[10]. **J.Chunfu**, **Y.Feng**. An Intrusion Detection Method Based on Hierarchical Hidden Markov Models. Wuhan University Journal of Natural Sciences. Vol. 12 No. 1 2007 135-128.
[11]. **S. Jha, K. Tan**, **R.A. Maxion**. Markov Chains,Classifiers and Intrusion Detection //Computer SecurityFoundations Workshop (CSFW), June 2001.
[12]. **L. E. Baum**,**T. Petrie**, "Statistical inference for probabilistic functions of finite state Markov chains," Annals of Mathematical Statistics, 37:1554-1563, 1966.
[13]. **S. Fine**, **Y. Singer**, **N. Tishby**. The Hierarchical Hidden Markov Model: Analysis and Applications[J]. MachineLearning, 1998, 32(1):41-62.
[14]. **Y. Ivanov**, **A.Bobick**.Recognition of Visual Activities and Interactions by Stochastic Parsing[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(8):852-872.
[15]. **M. Minsky**. Representation: Structuring Knowledge & Data in AI Programs. Association for the Advancement of Artificial Intelligence (AAAI). http://www.aaai.org/AITopics/pmwiki/pmwiki.php/AITopics/Representation
[16]. DARPA Agent Markup Language+Ontology Interface Layer. http://www.daml.org/2001/03/daml+oil-index 2001.
[17]. **J. Kopena**.DAMLJessKB. http://edge.mcs.drexel.edu/assemblies/software/damljesskb/articles/DAMLJessKB-2002.pdf, October 2002.
[18]. The Mitnick Case: How Bayes Could Have Helped. IFIP International Federation for Information Processing, 2005, Volume 194/2005, 91-104, DOI: 10.1007/0-387-31163-7_8.
[19]. **J. Zinky**, **R. Shapiro**, **S. Siracuse**, **T. Wright**. Experience with Dynamic Crosscutting in Cougaar. Lecture Notes in Computer Science, 2010, Volume 4803/2010, 595-612, DOI: 10.1007/978-3-540-76848-7_41.
[20]. **G. Isaza**, **A. Castillo**, **M. López**, **L. Castillo**. Towards Ontology-Based Intelligent Model for Intrusion Detection and Prevention. Advances in Soft Computing, 2009, Volume 63/2009, 109-116, DOI: 10.1007/978-3-642-04091-7_14.
[21]. **G. Antoniou**, **F. van Harmelen**. Web Ontology Language: OWL. Department of Computer Science, University of Crete.

**PavelOsipov**Mg.Sc.ing.Ph.D. student, Institute of Information Technology, Riga Technical University. He received his master's diploma from Transport and Telecommunications Institute, Riga. His research interests include web data mining, machine learning and knowledge extraction.Contact information: Institute of Information Technology, Riga Technical University, 1 Kalku Street, Riga, LV-1658, phone: +371 67089530, e-mail:pavels.osipovs@gmail.com

**ArkadyBorisov**is Professor of Computer Science in the Faculty of Computer Science and Information Technology at Riga Technical University. He holds a Doctor of Technical Sciences degree in Control in Technical Systems and the Dr. habil. sc. comp. degree. His research interests include fuzzy sets, fuzzy logic, computational intelligence and bioinformatics. He has 210 publications in the area. Contact information: Institute of Information Technology, Riga Technical University,1 Kalku Street, Riga, LV-1658, phone: +371 67089530, e-mail:arkadijs.borisovs@cs.rtu.lv.

**PāvelsOsipovs, ArkādijsBorisovs. Uzsabloniemnebalstītasmetodesanomālasuzvedībasatklāšanai**
Rakstāapskatītaspamatametodes, kastiekizmantotas, lainoteiktuvaiinformācijassistēmānavbijusiielaušanās. Šabloniemalternatīvumetožuaktualitātesaistītaarievērojamusarežģītībaspalielinājumumūsdienuinformācijasjomā, kasneļaujizmantotaprakstošāspieejasefektīvāielaušanās un anomālasuzvedībasgadījumuatklāšanā. Izmantojotjaunāsintelektuālāsmetodikas, kasizmantopašapmācībaselementus un mākslīgointelektu, kļūstiespējamsreālālaikāatklāt un reaģētuzjaunajiemielaušanāsveidiem.Izskatītapieeja, kastraktēnormālulietotājauzvedībukātroksni, bet anomālukāsignālu. Tādāgadījumāiriespējamsizmantotlabiizpētītossignālaapstrādes un filtrēšanasalgoritmus, kasvartiktpiemēroti, lainoteiktuielaušanosinformācijassistēmā.
Izskatītaarīvēlcitapieeja, kasizmantoslēptoshierarhiskos Markova modeļus, laiizveidotunormālalietotājauzvedībasmodeli. Šādumodeliiespējamsizmantottālākospētījumos, lainoteiktuanomālijaskatrānākamajālietotājadarbībāsistēmā.Tāpatapskatītaielaušanāspazīmjuontoloģijasveidošanaspieredze, kasļaujautomatizētzināšanuapmaiņustarpdažādamintelektuālāmdrošībassistēmāmnākotnē. Bez tam vēlizskatītaiespējamībalietotaģentupieeju, laiatklātuielaušanāsmēģinājumus, dots kopējaisaģentumijiedarbībasmodelis.Noslēgumā dots eksperimentālassistēmasizveidespiemērs, kasapvienodažādosmonitoringalīmeņosdažādaspieejas: aģentus, ontoloģijas, mākslīgosneironutīklus un klasifikācijasmetodes.

**Павел Осипов, АркадийБорисов.Методы обнаружения аномального поведения,не основанные на шаблонах**
В статье анализируются основные методы, используемые в задачах обнаружения вторжений в информационные системы. Актуальность методов обнаружения аномального поведения, не основанных на шаблонах,обусловлена значительным усложнением современной информационной среды, которое не позволяет и дальше эффективно использовать описательный подход к обнаружению вторжений и аномального поведения. С использованиемновых интеллектуальных методик с элементами самообучения и искусственного интеллекта появляется возможность в режиме реального времени отслеживать и реагировать на новые типы атак. Рассмотрен подход, трактующий нормальное поведение пользователя как шум, а аномальное - как сигнал. В этом случае появляется возможность использовать хорошо исследованные алгоритмы обработки и фильтрации сигнала применительно к предметной области обнаружения вторжений. Исследован подход, использующий скрытые иерархические модели Маркова для представления шаблона нормального поведения пользователя. Имея такую модель, возможно в дальнейшем использовать ё для анализа наличия аномальности для каждого последующего действия пользователя в системе. Также описан опыт построения онтологии признаков вторжений, которая позволит автоматизировать обмен знаниями между различными интеллектуальными системами безопасности в будущем.Кроме этого исследована возможность использования агентного подхода для обнаружения вторжений, предложена общая модель взаимодействия агентов.В заключение приведён пример создания экспериментальной системы, совмещающей на различном уровне мониторинга разные подходы: агентный, онтологии, искусственные нейронные сети и методы классификации.